面向 Cytoscape 平台的关联数据知识图谱概览抽取与可视化*

姜 赢 张 婧 朱玲萱
(北京师范大学珠海分校管理学院 珠海 519087)

摘要: 【**目的**】为更方便地查询和利用各个领域的海量关联数据，提出一种关联数据知识图谱概览的生成方法，使得用户在查询前就能了解关联数据访问点的内部数据结构。【**方法**】通过 SPARQL 查询关联数据所包含的领域知识关系，针对每一个知识关系构建知识图谱概览三元组并形成初步的知识图谱概览，再抽取每个知识分类的知识图谱概览三元组并合并到前者形成完整的知识图谱概览。【**结果**】研发 Cytoscape 插件实现此方法，并进一步提供知识图谱概览可视化功能。【**局限**】不能处理匿名节点等复杂知识分类抽取。【**结论**】在生物医学领域分别进行单点抽取、关联"桥"和关联"包含"三项测试，测试结果表明该方法抽取速度快而稳定，抽取结果的查全率高，且不需要网络爬虫或额外的索引工作。

关键词: 关联数据 知识图谱概览 SPARQL Cytoscape
分类号: TP393

1 引 言

关联数据[1](Linked Data)提出的目的是构建计算机能理解的语义数据网络，而不仅仅是人能读懂的文档网络，以便于在此之上构建更智能的应用。关联数据主要通过资源描述框架 RDF[2](Resource Description Framework)格式来表示，它将一个资源描述成一组三元组(主体，谓语，客体)。SPARQL[3]是 W3C 推荐标准，提供对 Web 上或 RDF 存储中的内容进行查询和处理的语言和协议。每个关联数据都提供 SPARQL 访问点[4](SPARQL Endpoint)，它是一种 HTTP 绑定协议，用于通过 HTTP 进行 SPARQL 查询，并返回相应数据。例如，Linked Open Data[1]项目的宗旨在于号召将现有数据发布成关联数据，并将不同数据源互联起来。截至 2014 年，1 014 个包含数十亿 RDF 三元组的数据集在 Linked Open Data 项目中建立了关联[5]，涵盖医疗卫生、电子商务、生物化学、国防军事、人文历史等各个领域。

尽管关联数据和 RDF 本体形式化的优点在于能够使异构分布式大数据进行无缝整合，但是这种形式化需要一种格式将数据从预定义模式的枷锁中解脱出来。使用这种格式定义的数据为查询带来了挑战，因为用户无法依赖事先获得的数据映射进行 SPARQL 查询。也就是说，虽然关联数据可以通过 SPARQL 查询终端进行查询，但是对于某个特定领域的数据，用户在提交查询请求前必须事先知道此关联数据的内部数据结构，也就是知识图谱概览[6](Knowledge Graph Schema)。知识图谱概览描述关联数据包含哪些知识分类(Class)、知识关系(Object Property)，以及知识关系如何关联知识点(Instance)组成的知识分类[6]。在此基础之上，用户才能写出 SPARQL 语句进行查询，否则将无法针对海量数据进行查询。这就如同查询数据库之前，首先要了解数据库的表结构(Schema)才行。因此，本文提出一种快速有效抽取关联数据知识图谱概

通讯作者：姜赢，ORCID：0000-0002-2596-5242，E-mail：jpz6311whu@bnuz.edu.cn。
*本文系广东省高等学校优秀青年教师培养计划项目"面向大数据的生物通路本体知识图谱可视化研究"(项目编号：YQ2015239)和广东省自然科学基金项目"基于本体推理演化的财经大数据分析与预测研究"(项目编号：2016A030313386)的研究成果之一。

览并进行可视化的方法, 并在 Cytoscape[7]平台上以插件形式进行实现, 力求降低用户查询和使用海量关联数据的门槛。

2 研究现状介绍

知识图谱概览的提取和可视化对知识工程意义重大, 目前主要用于辅助感知、构建和调试知识图谱的结构和内容[8]。以 2007 年关联数据的提出为转折点, 国内外相关算法和工具的研究经历了以下两个阶段。

2.1 知识图谱概览的可视化

由于服务对象多为本体设计和构建人员, 早期知识图谱概览的可视化工具, 大多数以本体编辑器的插件形式提供给用户。例如, 基于 Protégé[9]本体编辑器的可视化插件包括最早在 2003 年研发的 TGVizTab[10]、Protégé 默认的内置插件 OntoGraf [11]和 Protégé-OWL[12]插件 NavigOWL[13]等。还有基于 Neon-Toolkit[14]本体编辑器的可视化插件 KC-Viz[15]等。另外, GrOWL[16]等少量工具是以独立的 Java 桌面程序形式(非插件)提供给用户。以上这些可视化工具有如下特征。

(1) 面向有向图的可视化思路

大多数可视化工具的功能是将知识图谱概览中的知识分类和知识关系映射为节点(Node)和边(Edge)以形成有向图, 再通过某种布局(Layout)算法(例如, Spring Layout[10]和 Power-Law Graph[13])在特定显示区域内展示静态或动态的知识图谱概览示意图。虽然还有个别其他思路(例如, 将知识图谱概览转换为 UML 图例的 OWLGrEd[17]和 OntoViz[18]), 但有向图是知识图谱概览可视化的主流思路。

(2) 面向本地数据的可视化机制

为将知识图谱概览进行可视化, 用户必须将知识图谱数据导入到这些可视化工具中。这要求数据存储在本地客户端, 而且用户要有完整的数据。也就是说, 除非用户是知识图谱的作者或是拥有完整数据读取权限的使用者, 否则一般的用户无法进行知识图谱概览可视化操作。虽然也有一些如 WebVOWL[19]这样的本体编辑器提供基于 Web 的可视化操作入口, 但是操作之前还需要将完整的数据上传到服务器。总之, 现有的可视化工具只能处理完整的本地数据, 无法处理以 SPARQL 访问点查询为基础的关联数据知识图谱可视化。

(3) 海量数据的可视化局限性

现有的本体编辑器往往限制显示在图面板中的节点数少于 10 000, 这样避免了海量数据的可视化问题[20]。例如, OntSphere[21]工具分析报告指出稍微超过 1 000 节点会导致其显示节点或标签时出现重叠问题; TGVizTab 和 OntoViz 在节点数超过 300 之后便无法进行可视化。NavigOWL 在 Power-Law Graph 布局算法上进行改进, 在节点数为 10 000 至 100 000 的数据集上进行对比测试, 结果显示其性能优于传统的 Spring Layout 可视化算法。即便如此, 包括 NavigOWL 在内的这些工具也难以可视化数量级在百万、千万以上的关联数据知识图谱。

2.2 知识图谱概览的抽取

自 2007 年关联数据知识图谱提出以来, 其分布式、海量数据的内在特征使得早期的可视化工具无法满足要求。越来越多不同领域的关联数据被建立起来, 其推广和使用亟需向用户提供关联数据知识图谱概览的可视化方法和工具。这个矛盾使得人们从知识图谱概览可视化的研究转向知识图谱概览抽取的研究。研究方向转变的原因在于: 关联数据虽然是海量的且难以直接进行可视化, 但是绝大多数的数据是具体的事实数据, 而其中知识图谱概览相关数据并不多。这正如同数据库记录条数可以有上千万条, 但是表结构和字段个数并不是海量的。如果能够在对关联数据进行可视化之前, 先将它的知识图谱概览抽取出来, 可以大大降低可视化内容的数量级。基于以上思路, 近年来关联数据研究人员研发出知识图谱概览抽取的方法和工具, 主要有以下三种。

(1) A 方法: 基于网络爬虫的关联数据索引方法

2011 年, Semantic Web Challenge(Billion Triples Track)获奖者 Konrath 等研发出 SchemEX 工具[6], 可以利用爬虫对海量关联数据建立实时的索引, 并提供线性时间复杂度下的知识图谱概览抽取功能。SchemEX 具有面向 RDF 数据流的动态索引机制, 这使得抽取与具体的爬虫方法松散耦合。该工具缺点是: 需要通过网络爬虫获取、扫描关联数据的全部数据, 而且需要做额外的数据索引; 如果关联数据不允许爬取, 则无法处理。

(2) B 方法: 以知识分类为切入点的 SPARQL 查询方法

主要有两种: B1 方法通过 rdf:type[22]和 rdfs:

subclassOf[22]等获得 RDF Schema[22](RDFS)级别的知识分类“层级(Hierarchy)关系概览”; B2 方法通过 owl:domain[23]、owl:range[23]、owl:Class 等获得 Web Ontology Language[24](OWL)级别的知识分类“关联(Relation)关系概览”。这些方法只考虑显性的知识图谱概览, 也就是关联数据明确定义知识分类关系概览的情况。但是, 根据 Gottron 等对大量关联数据知识图谱概览的计量研究发现, 这种方法只能获得 63.5%到 88.1%的知识图谱概览信息[25]。原因在于: 关联数据没有明确定义 rdf:typerdfs:Class 和 owl:Class 的情况是常有的现象, 而隐性知识图谱概览几乎在每个关联数据中都会出现。

(3) C 方法: 基于 RDF 图摘要的替代方法。

Zneika 等[26]提出基于 top-k 近似拟合的 RDF 图摘要生成方法, 可以将海量关联数据转换成描述知识库内容的知识图谱概览。此方法缺点在于转换获得的知识图谱概览只是关联数据内容和结构的“近似”, 存在一定误差。

与上述三种方法不同, 本文提出以知识关系为切入点、完全使用 SPARQL 查询的知识图谱概览抽取实现方法。该方法在抽取步骤中借鉴融合 B 方法的部分思路。在可视化思路上, 仍然采纳以有向图作为可视化结果的思路。本文方法抽取速度快而稳定, 抽取结果查全率高, 而且不需要网络爬虫或额外的索引工作, 也避免关联数据未能明确定义 rdf:type、rdfs:Class 和 owl:Class[23]的遗漏情况。

3 研究思路与框架

知识图谱概览抽取思路如图 1 所示。本文提出的知识图谱概览抽取方法主要包含 5 个步骤:

(1) 查询关联数据所包含的知识关系集合 P。SPARQL 查询语句为:

```
SELECT distinct ?p WHERE { ?s ?p ?o .}
```

(2) 过滤掉集合 P 中以 rdf(http://www.w3.org/1999/02/22-rdf-syntax-ns), rdfs(http://www.w3.org/2000/01/rdf-schema#)和 owl(http://www.w3.org/2002/07/owl#)为命名空间的与领域知识无关的知识关系, 得到集合 P’。具体来说, 对于知识关系集合 P 中的每一个知识关系 p, 抽取它的命名空间 n; 如果命名空间 n 是 rdf, rdfs 或 owl, 则将 p 纳入到待过滤的知识关系集合 Q; 最后将 Q 中的每个元素从 P 中删除, 得到集合 P’。

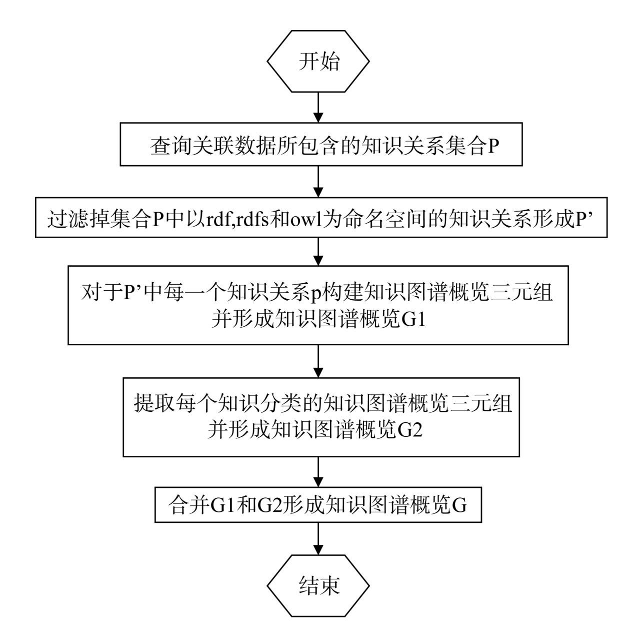


图 1 知识图谱概览抽取思路

(3) 构建集合 P’中每一个知识关系 p 的知识谱图概览 G1。

以主体和客体为切入点, 构建知识关系 p 的知识图谱概览 G1: 查询以知识关系 p 为谓语的三元组的主体 s 和客体 o, 分别查询主体 s 和客体 o 的知识分类集合 C1 和 C2; 如果 C1 和 C2 都不为空, 则对于每一个(c1, c2)组合(其中 c1∈C1, c2∈C2), 构建以 c1 为主体、p 为谓语和 c2 为客体的知识图谱概览三元组, 并纳入到知识图谱概览 G1 中。SPARQL 查询语句为:

```
CONSTRUCT { ?class1 <" + p + "> ?class2 } WHERE { ?s <" + p + "> ?o . ?s a ?class1 . ?o a ?class2. }"
```

以谓语为切入点, 构建知识关系 p 的知识图谱概览 G1: 查询以知识关系 p 为主语、以 rdfs: domain[25]为谓语的三元组客体 c1; 查询以知识关系 p 为主语、以 rdfs:range[25]为谓语的三元组客体 c2; 如果 c1 和 c2 都能查询到, 则构建以 c1 为主体、p 为谓语和 c2 为客体的知识图谱概览三元组, 并纳入到知识图谱概览 G1 中。

(4) 查询描述知识分类之间直接关系的知识图谱三元组, 并纳入到知识图谱概览 G2 中。具体来说, 查询以 rdfs:subClassOf、owl:equivalentClass[23]、owl: complementOf[23]或 owl: disjointWith[23]为谓语的三元组, 如果此三元组的主体和客体都不是匿名节点, 则将这个三元组纳入到知识图谱概览 G2 中。

(5) 将G1与G2合并到G, G就是最终的知识图谱概览。具体来说, 去掉知识图谱概览G1中重复的知识图谱概览三元组, 形成知识图谱概览G1'; 去掉知识图谱概览G2中重复的知识图谱概览三元组, 形成知识图谱概览G2'; 将G1'和G2'合并成知识图谱概览G'; 去掉知识图谱概览G'中重复的知识图谱概览三元组, 最终形成知识图谱概览G。

4 实验过程与结果

4.1 实验对象

由于近年来生物医学领域出现了一大批可以查询的结构化生物医学关联数据, 因此选取生物医学领域关联数据作为实验对象。对于生物医学研究人员来说, 将这些海量关联数据用于日常研究工作的门槛非常高, 特别在生物医学领域的用户交互接口(SPARQL访问点)和SPARQL查询整合方面的问题较多。这会使生物医学关联数据不能得到充分利用, 很多关联数据内所包含的知识分类、知识点与知识关系并不能被生物医学人员发现和利用。同时, SPARQL的语法问题与关联数据的复杂性问题也提高了生物医学研究人员的使用门槛。

在这样的背景之下, 选取几个最常用的、免费开放的生物通路本体关联数据作为测试数据集, 力求辅助生物医学研究人员完成对生物医学关联数据的查询、检索、解析与重新组织。具体来说, 实验对象包括: 多伦多大学Bader实验室研发的Pathway Common (PC)[27]、SRI International公司提供开放学术研究License的BioCyc[28]、开源的Reactome[29]、伦敦大学学院研发的基因本体HGNC(Human Gene Nomenclature Database)[30]和欧洲生物信息研究所研发的BioModel[31]。

4.2 实验工具

Cytoscape是NBRB开源组织开发的一个专注于开源网络可视化分析的平台, 其核心是提供基础的功能布局和查询网络, 并依据基本数据形成可视化网络。它最先应用于生物学领域, 一般用于整合复杂分子间的相互作用网络和相关分子的状态信息, 也广泛用于可视化蛋白质、DNA等数据库。

在平台功能上, Cytoscape可使用不同的可视化样式显示生物分子相互作用网络。能够在两个维度上布局网络, 且有多种布局算法可供选择(包括环状和弹簧嵌入式布局); 可自由浏览、放大、缩小或平移网络, 并提供使用鸟瞰的形式导航大型网络(100 000节点和边)且具有高效的渲染引擎; 可以对网络图的节点、连线、边框等进行注释, 并能够自定义标签和颜色; 还可以轻松组织和管理多个网络, 并支持将网络结构保存在一个会话文件中留存备用。但是, Cytoscape平台本身没有关联数据可视化功能, 也没有知识图谱概览抽取功能。

笔者研发了基于Cytoscape平台的关联知识图谱概览抽取插件, 并以此作为实验工具分别进行单点抽取、关联"桥"和关联"包含"三项实验。

4.3 单个关联数据访问点抽取实验

(1) 不同数据实验对象的对比实验

图2是生物通路本体Pathway Commons[27]关联数据访问点的知识图谱概览可视化图。其中圆形节点表示的是知识分类, 而节点之间的箭头是知识关联。箭头的指向是从主语知识分类到宾语知识分类。知识分类和知识关联都有文字Tooltip显示它们的URI。

实验结果如表1所示, 大部分SPARQL访问点的知识图谱概览抽取时间在10分钟以内, 可以达到实用层次。用户抽取完成之后, 可以将可视化结果进行保存, 下次使用时无需重新抽取。

(2) 不同抽取方法的对比实验

以MeSH(Medical Subject Headings)[28]关联数据为实验对象, 将本文提出的抽取方法与B方法(包含B1和B2两个方法)进行对比实验。由于A方法和C方法的作者未提供相关抽取工具下载, 所以只选取B方法进行分析。

经过初步分析, MeSH关联数据尽管包含三元组总数达到60多万个, 但只包含28个知识分类和152个知识关系。它以自身的meshv(http://id.nlm.nih.gov/mesh/vocab#)命名空间词表为主, 并引入部分VirtRDF[29]、DC[30]和FOAF[31]词表。如表2所示, 知识关系中标记rdfs:domain和rdfs:range的分别只有6个和10个; 知识分类中标记owl:Class的只有16个且均属于meshv命名空间词表, 而明确标记rdfs:Class的为0个。如表3所示, 只有8个标记rdfs:subClassOf的知识分类; 除去owl:Thing[23]就只有meshv:broaderQualifier[28]和meshv:Qualifier[28]这一组知识分类明确标记子类、父类之间的关系。

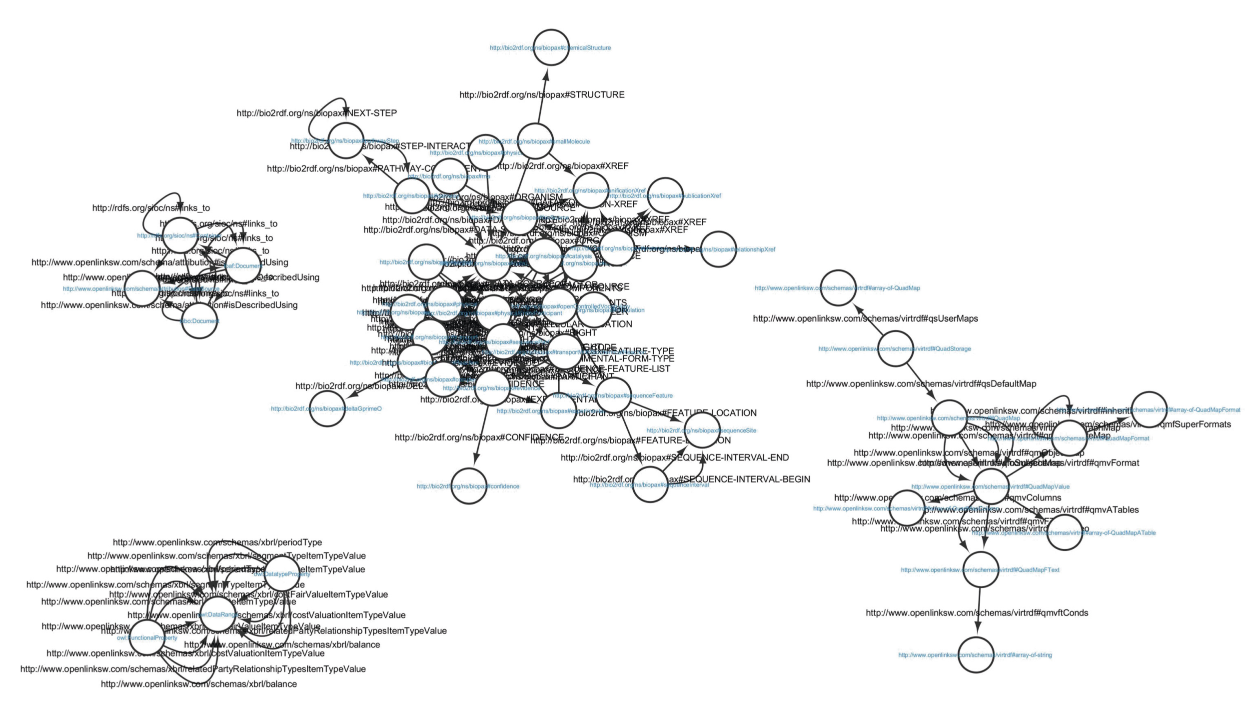


图 2　Pathway Commons 的关联数据访问点的知识图谱概览抽取图

表 1　单个关联数据访问点的知识图谱概览抽取实验结果表

关联数据 SPARQL 访问点	RDF 三元组个数	抽取时间(分钟)
Pathway Commons	27 623 683	8.16
BioCyc	18 532 342	9.57
MeSH	654 198	10.86
Reactome	2 980 230	6.45

表 2　MeSH 关联数据中标记 rdfs:domain 和 rdfs:range 的知识关系

rdfs:domain 知识分类	知识关系	rdfs:range 知识分类
meshv:TreeNumber	meshv:parentTreeNumber	meshv:TreeNumber
—	meshv:treeNumber	meshv:TreeNumber
meshv:Concept	meshv:broaderConcept	meshv:Concept
meshv:Concept	meshv:narrowerConcept	meshv:Concept
meshv:Concept	meshv:relatedConcept	meshv:Concept
meshv:Descriptor	meshv:broaderDescriptor	meshv:Descriptor
—	meshv:hasDescriptor	meshv:Descriptor
—	meshv:allowableQualifier	meshv:Qualifier
—	meshv:hasQualifier	meshv:Qualifier
meshv:Qualifier	meshv:broaderQualifier	meshv:Qualifier

表 3　MeSH 关联数据中标记 rdfs:subClassOf 的知识分类

子类(知识分类)	父类(知识分类)
meshv:TreeNumber	owl:Thing
meshv:Concept	owl:Thing
meshv:Descriptor	owl:Thing
meshv:DescriptorQualifierPair	owl:Thing
meshv:SupplementaryConceptRecord	owl:Thing
meshv:Qualifier	owl:Thing
meshv:Term	owl:Thing
meshv:broaderQualifier	meshv:Qualifier

如表 4 所示，在 MeSH 关联数据所有应该被抽取的35个知识图谱概览三元组中，B方法只抽取出14个，查全率只有 40%；本文提出的方法查全率达到94.28%。如图 3 所示，B 方法未能抽取的知识图谱概览三元组主要来自 VirtRDF 和 MeSH 中未能明确标记的知识分类；本文方法可以将它们抽取出来，主要原因是此方法遵循自底向上思路，即通过知识关系相对于底层知识点的关联来反推上层知识分类之间的关联，从而避免 rdfs:domain 和 rdfs:range 等知识分类标记缺失的问题。另外，本文方法仍有两个知识图谱概览三元组未能抽取出来，是因为 MeSH 在定义 owl:Thing

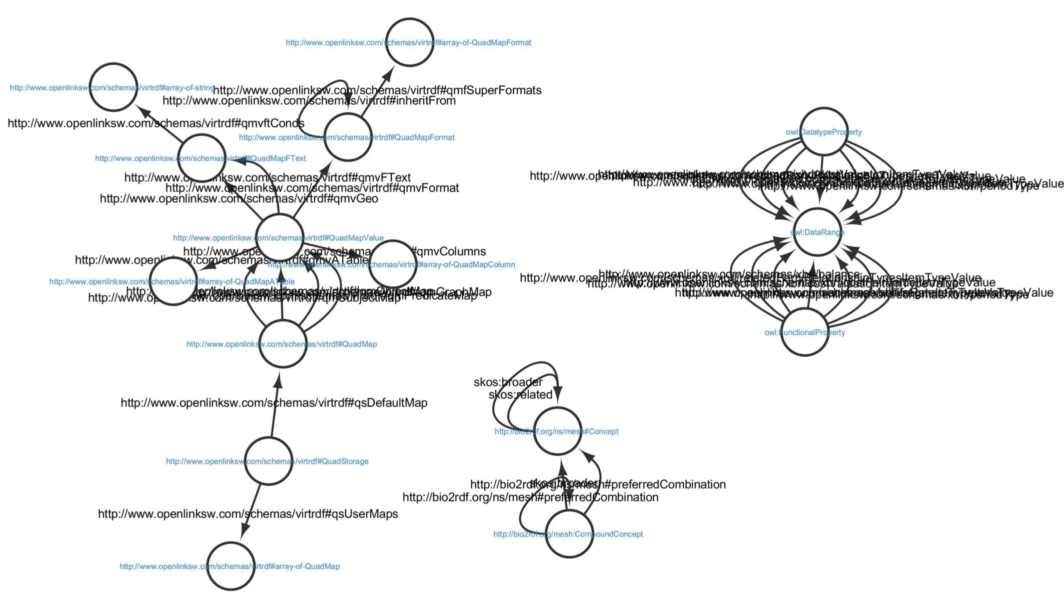


图 3　B 方法未能抽取的知识图谱概览三元组

表 4　不同抽取方法的查全率对比表

对比项目	B1 方法	B2 方法	B 方法 (B1+B2)	本文方法
抽取知识图谱概览三元组数量	8	6	14	33
抽取查全率	22.86%	17.14%	40.00%	94.28%

时使用了两个基于匿名节点的复杂知识分类，导致本文方法的查全率未达到 100%。

4.4　多个关联数据访问点的抽取实验

关联数据之间并不是孤立的。不同数据源可以来自一个组织内部的不同系统，也可以来自不同组织的不同系统。数据源的内容、存储地点以及存储方式都可以完全不同，但它们仍可能存在关联。笔者研发的 Cytoscape 插件还能够抽取多个关联数据访问点的知识图谱概览并展示它们之间的关联，主要包含以下两种情况：关联“包含”和关联“桥”。

(1) 关联“包含”

如表 5 所示，笔者抽取了两个生物通路关联数据(HGNC[32]和 MeSH)的知识图谱概览，运行时间约为 3 分钟。图 4 的可视化结果中用不同的形状来表示来自不同 SPARQL 访问点的知识分类节点。但是，可视化结果中并没有出现表示来自 MeSH 数据源的知识分类(三角形节点)，因为 HGNC 数据源的知识图谱概览包含 MeSH 数据源的所有内容，被包含的部分数据也显示为圆形。

表 5　HGNC 和 BioCyc 关联“包含”的知识图谱概览抽取实验结果

关联数据 SPARQL 访问点	RDF 三元组个数	图示节点形状
HGNC	922 523	圆形
MeSH	654 198	三角形
关联“包含”	-	圆形

(2) 关联“桥”

在图 5 的可视化结果中，Pathway Commons、Linkedspl 和 BioModel 的关联“桥”是三角形和方形之间的圆形节点，这些圆形节点就是三个关联数据同时包含的知识分类。图 5 中右上方的放大子图也显示了关联“桥”是知识分类“http://www.biopax.org/release/biopax-level3.owl#Pathway”。用户还可以通过每个知识分类圆形节点的“endpoint”属性来查看其所属的 SPARQL 访问点。当用户获得这个知识图谱概览可视化图之后，还可以通过这个关联“桥”进一步实现跨数据源的关联查询。这三个关联数据的数据信息如表 6 所示。

表 6　关联“桥”的知识图谱概览抽取实验结果

关联数据 SPARQL 访问点	RDF 三元组个数	图示节点形状
BioModel	2 380 009	三角形
Pathway Commons	27 623 683	方形
Linkedspl	2 174 579	菱形
关联“桥”	-	圆形

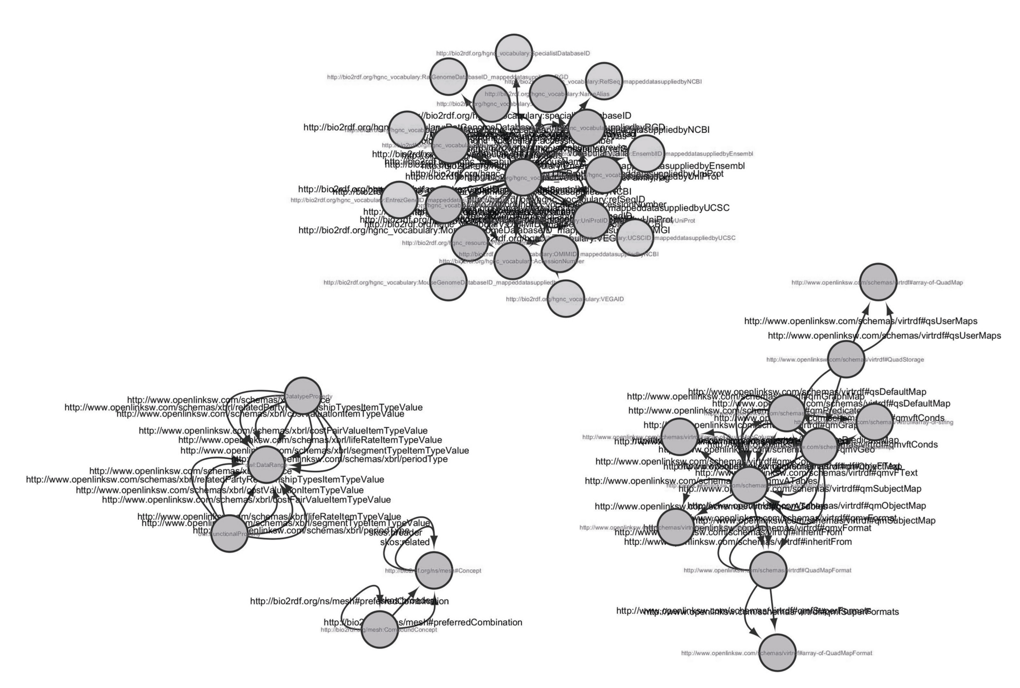


图 4　HGNC 和 BioCyc 关联“包含”知识图谱概览抽取可视化图

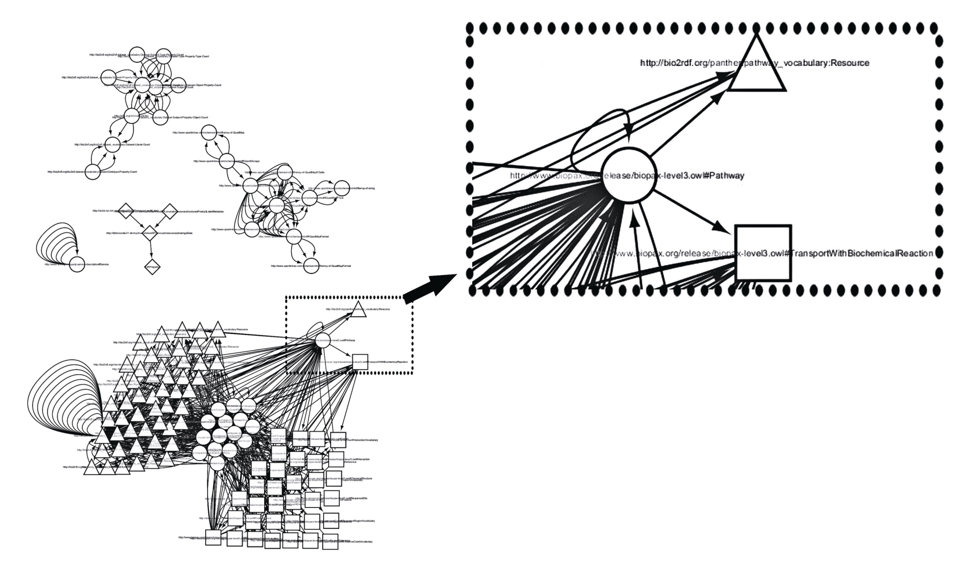


图 5　关联“桥”知识图谱概览抽取可视化图

5　结　语

本文关联数据知识图谱概览抽取方法以 Cytoscape 平台为依托，通过分步抽取方式整合各类知识图谱概览。这是一种自下而上的抽取思路。通过生物医学领域关联数据实验测试表明，该方法以知识关系为切入点、完全利用 SPARQL 查询来实现抽取；该方法抽取速度快而稳定、抽取结果的查全率高，且不需要网络爬虫或额外的索引工作。未来工作主要包括：研究以匿名节点为基础的复杂知识分类概览抽取算法，例如 owl:unionOf [23]、owl:allValuesFrom[23]等，并整合到关联数据知识图谱概览抽取方法中；完善插件功能，并提供 OWL2[12]知识图谱概览的可视化功能。

参考文献：

[1]　Bizer C, Heath T, Berners-Lee T. Linked Data—The Story So Far[J]. International Journal on Semantic Web & Information Systems, 2009, 5(3): 1-22.

[2] Klyne G, Carroll J J, McBride B. RDF 1.1 Concepts and Abstract Syntax [EB/OL]. (2014-02-25). [2017-01-25]. https://www.w3.org/TR/rdf11-concepts/.

[3] Harris S, Seaborne A. SPARQL 1.1 Query Language [EB/OL]. (2013-03-21). [2017-01-25]. https://www.w3.org/TR/sparql11-query/.

[4] Feigenbaum L, Williams G T, Clark K G, et al. SPARQL 1.1 Protocol [EB/OL]. (2013-03-21). [2017-01-25]. https://www.w3.org/TR/sparql11-protocol/.

[5] Schmachtenberg M, Bizer C, Paulheim H. Adoption of the Linked Data Best Practices in Different Topical Domains[C]//Proceedings of the 13th International Semantic Web Conference (ISWC 2014), Seattle, USA. Germany: Springer, 2014.

[6] Konrath M, Gottron T, Scherp A. Schemex-Web-Scale Indexed Schema Extraction of Linked Open Data[C]//Proceedings of the 13th International Semantic Web Conference (ISWC 2011) Submission to the Billion Triple Track, Bonn, Germany. Springer, 2011:52-58.

[7] Shannon P, Markiel A, Ozier O, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks [J]. Genome Research, 2003, 13(11): 2498-2504.

[8] Lanzenberger M, Sampson J, Rester M. Visualization in Ontology Tools[C]//Proceedings of the 2nd International Conference on Complex, Intelligent and Software Intensive Systems (CISIS 2009), Burgos, Spain. New York, USA: IEEE, 2009: 705-711.

[9] Gennari J H, Musen M A, Fergerson R W, et al. The Evolution of Protégé: An Environment for Knowledge-based Systems Development[J]. International Journal of Human-Computer Studies, 2003, 58(1): 89-123.

[10] Alani H. TGVizTab: An Ontology Visualisation Extension for Protégé [C]//Proceedings of the 2nd International Conference on Knowledge Capture (K-Cap'03), Florida, USA. ACM, 2003.

[11] Falconer S. OntoGraf [EB/OL]. [2017-01-25]. http://protegewiki.stanford.edu/wiki/OntoGraf.

[12] Knublauch H, Fergerson R W, Noy N F, et al. The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications[C]//Proceedings of the 3rd International Semantic Web Conference (ISWC 2004), Karlsruhe, Germany. Springer, 2004: 229-243.

[13] Hussain A, Latif K, Rextin A T, et al. Scalable Visualization of Semantic Nets Using Power-Law Graphs [J]. Applied Mathematics & Information Sciences, 2014, 8(1): 355-367.

[14] Haase P, Lewen H, Studer R, et al. The Neon Ontology Engineering Toolkit[C]//Proccedings of the 17th International World Wide Web Conference (WWW 2008), Beijing, China. ACM, 2008.

[15] Motta E, Mulholland P, Peroni S, et al. A Novel Approach to Visualizing and Navigating Ontologies[C]//Proceedings of the 10th International Semantic Web Conference (ISWC 2011), Bonn, Germany. Springer, 2011: 470-486.

[16] Krivov S, Villa F, Williams R, et al. On Visualization of OWL Ontologies[A]// Baker C J, Cheung K H. Semantic Web[M]. Springer Berlin Heidelberg, 2007: 205-221.

[17] Bārzdiņš J, Bārzdiņš G, Čerāns K, et al. UML Style Graphical Notation and Editor for OWL 2[C]//Proceedings of the 9th International Conference on Business Informatics Research (BIR 2010), Rostock, Germany. Springer, 2010: 102-114.

[18] Sintek M. OntoViz [EB/OL]. [2017-01-25]. http://protegewiki.stanford.edu/wiki/OntoViz.

[19] Lohmann S, Link V, Marbach E, et al. WebVOWL: Web-based Visualization of Ontologies[C]//Proceedings of the International Conference on Knowledge Engineering and Knowledge Management (EKAW 2014), Linköping, Sweden. Springer, 2014: 154-158.

[20] Katifori A, Halatsis C, Lepouras G, et al. Ontology Visualization Methods—A Survey[J]. ACM Computing Surveys, 2007, 39(4): Article No. 10.

[21] Bosca A, Bomino D, Pellegrino P. OntoSphere: More than a 3D Ontology Visualization Tool[C]// Proceedings of the 2nd Italian Semantic Web Workshop (SWAP 2005), Trento, Italy. 2005.

[22] Brickley D, Guha R V. RDF Schema 1.1[EB/OL]. (2013-03-21). [2017-01-25]. http://www.w3.org/TR/rdf-schema/.

[23] Bechhofer S, Harmelen F V, Hendler J, et al. OWL Web Ontology Language Reference [EB/OL]. (2004-02-10). [2017-01-25]. https://www.w3.org/TR/owl-ref/.

[24] Grau B C, Horrocks I, Motik B, et al. OWL 2: The Next Step for OWL[J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2008,6(4): 309-322.

[25] Gottron T, Knauf M, Scheglmann S, et al. A Systematic Investigation of Explicit and Implicit Schema Information on the Linked Open Data Cloud[C]//Proceedings of the 10th European Semantic Web Conference (ESWC 2013), Montpellier, France. Springer, 2013: 228-242.

[26] Zneika M, Lucchese C, Vodislav D, et al. RDF Graph Summarization Based on Approximate Patterns[C]//

Proceedings of the International Workshop on Information Search, Integration, and Personalization (ISIP 2015), Grand Forks, USA. Springer, 2015: 69-87.

[27] Cerami E G, Gross B E, Demir E, et al. Pathway Commons, A Web Resource for Biological Pathway Data[J]. Nucleic Acids Research, 2011, 39(Database Issue): D685-D690.

[28] Caspi R, Billington R, Ferrer L, et al. The MetaCyc Database of Metabolic Pathways and Enzymes and the BioCyc Collection of Pathway/Genome Databases[J]. Nucleic Acids Research, 2014, 42(Database Issue): D459-D471.

[29] Joshi-Tope G, Gillespie M, Vastrik I, et al. Reactome: A Knowledgebase of Biological Pathways[J]. Nucleic Acids Research, 2005, 33(Database Issue): D428-D432.

[30] Bruford E A, Lush M J, Wright M W, et al. The HGNC Database in 2008: A Resource for the Human Genome[J]. Nucleic Acids Research, 2008, 36(Database Issue): D445-D448.

[31] Novère N L, Bornstein B, Broicher A, et al. BioModels Database: A Free, Centralized Database of Curated, Published, Quantitative Kinetic Models of Biochemical and Cellular Systems[J]. Nucleic Acids Research, 2006, 34(Database Issue): D689-D691.

[32] Erling O, Mikhailov I. RDF Support in the Virtuoso DBMS[A]//Networked Knowledge-Networked Media[M]. Springer Berlin Heidelberg, 2009: 7-24.

[33] Weibel S. The Dublin Core: A Simple Content Description Model for Electronic Resources[J]. Bulletin of the Association for Information Science and Technology, 1997, 24(1): 9-11.

[34] Brickley D, Miller L. FOAF Vocabulary Specification 0.91[EB/OL]. (2007-11-02). [2017-01-25]. http://xmlns.com/foaf/spec/20071002.html.

作者贡献声明:

姜赢: 提出研究思路, 设计研究方案, 论文撰写与修订;
张婧: 进行实验, 采集、清洗和分析数据;
朱玲萱: 研发模型系统。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: jpz6311whu@bnuz.edu.cn。
[1] 姜赢. src.rar. Cytoscape 插件程序源代码.

收稿日期: 2017-01-18
收修改稿日期: 2017-02-14

Extracting and Visualizing Knowledge Graph Schema from Linked Data with Cytoscape Platform

Jiang Ying　Zhang Jing　Zhu Lingxuan
(School of Management, Beijing Normal University, Zhuhai, Zhuhai 519087, China)

Abstract: [**Objective**] This paper proposes a new method to generate knowledge graph schema, aiming to help us understand the data structure before submitting a query, and improve the performance of linked data retrieval. [**Methods**] First, we searched knowledge relations of the linked data through SPARQL. Second, we constructed knowledge graph schema triples for each identified relation. Finally, we extracted graphs schema triples from every knowledge class and merged them with those of the relations. [**Results**] A Cytoscape plugin was developed based on the proposed method to visualize the knowledge graph schema. [**Limitations**] Our method could not extract knowledge from complex classtification, such as anonymous nodes. [**Conclusions**] The proposed method was examined with biomedical data for single, inclusive, and bridge extractions. It could retrieve information effectively, and does not need additional crawling and index efforts.

Keywords: Linked Data　Knowledge Graph Schema　SPARQL　Cytoscape